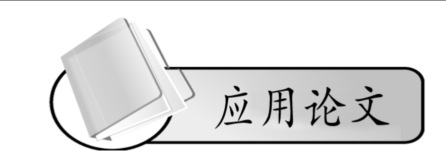


构建面向 WARC 文档的全文索引系统

胡吉颖　吴振新　谢　靖　张智雄

(中国科学院文献情报中心　北京 100190)

摘要:【**目的**】开发网络信息存档 WARC 文件的解析与索引系统，充分挖掘科技网站存档资源价值。【**应用背景**】在网络资源采集存档领域，WARC 文件格式获得了广泛的应用。随着网络信息的多样化，已有的 WARC 文件索引工具越来越难以满足用户多样性的查询需求。【**方法**】采用模块化方案解析 WARC 文件。分析比较常用的索引工具，选择 Solr 平台开发全文索引系统。【**结果**】实现对 WARC 文件基于内容的检索访问服务，并在 WARC 的索引中增加了学科分类、资源类型和存档时间等分面检索内容，从多维度对 WARC 文件内容进行揭示。【**结论**】向用户提供了丰富的科技网站存档数据信息，提高了用户检索访问效率。

关键词: 网络存档　WARC 文件　模块化解析　Solr 索引

分类号: G352

1　引　言

随着互联网在社会生活中的深度扩展，网络信息的更新速度不断加快，网络资源的生命周期变得更加短暂。采集并保存具有价值的 Web 资源信息，以满足当前和未来的访问和使用，已成为国内外许多保存机构的重要工作，一些重要在线资源的存档，如科研和商业信息，已经上升到国家战略层面。

国际网络保存联盟(IIPC)[1]自成立以来，一直致力于推动全球网络存档(Web Archive)活动，其成员包括全球 40 多个国家图书馆、档案馆等，在网络信息保存领域占有十分重要的地位。为了更好地支持采集和访问，以及存档机构之间的交流，IIPC 提出了标准化的网络文件存档格式 WARC(Web Archive)，并于 2009 年成为 ISO 国际标准[2]。WARC 文件格式具有记录信息量大、支持扩展和压缩以及易于管理的特点，已在许多 Web Archive 项目中得到广泛应用[3]。

随着参与保存的科研机构数量增多，存档资源数据量的迅速上升，存档资源信息的利用与共享成为 Web Archive 活动中越来越重要的方面。如何提升用户体验，满足用户多样性的查询需求；如何利用先进的技术有效重用存档资源，充分发挥存档资源的价值，这些都成为存档项目开发面临的挑战。WARC 文件解析和索引技术的提高，成为解决这些问题的关键环节。本文将详细介绍中国科学院文献情报中心 Web Archive 团队在国际重要科研机构网络资源存档实践中，如何有效地实施 WARC 的解析及构建全文索引系统。

2　WARC 索引工具现状及需求分析

网络存档内容不仅数据量巨大，而且随着时间在非线性地动态增加。为了实现 WARC 存档文件的检索和访问，对 WARC 的解析与索引显得尤为重要，最常用的开源软件工具包括 Wayback[4]、NutchWAX[5]和 WERA[6]等。

Wayback 是互联网档案馆(Internet Archive, IA)提供的一个基于 URL 的索引与访问工具，可以重现时间轴上不同时间点存档的页面，但不支持对文本内容的检索访问，因而难以满足用户多样化的搜索需求。NutchWAX[7]可以实现基于内容的全文索引和检索，并在许多国家的网络存档项目中得到应用。目前，Nutch 开发团队战略方向发生了变化[8]，已很少提供

通讯作者：吴振新，ORCID：0000-0003-4966-1961，E-mail：wuzx@mail.las.ac.cn。

NutchWAX 的后续维护和更新，而专注于网络爬虫工具的开发。此外，NutchWAX 依赖于 Hadoop 的文件系统建立索引，不具备在本地建立索引的能力；功能扩展比较困难，需要修改大量的代码才能实现。因此，NutchWAX 已不再是一个合适的提供全文检索的平台。WERA 也是一款比较常用的网络存档内容查询与浏览工具，它由 IA 和挪威国家图书馆共同开发，既可以提供类似于 Wayback 的 URL 检索功能，也具备全文检索的能力。但是，开发人员采用 NutchWAX 作为其搜索引擎的内核，因而 WERA 也没有从根本上解决 NutchWAX 所面临的局限性。Solr[9]是应用十分广泛的开源企业级搜索软件，具备全文索引和分面检索功能。IA 在对比 NutchWAX 和 Solr 性能后，发现在保存的文档数量急剧增加时，NutchWAX 的检索性能明显低于 Solr，因而推荐采用 Solr 作为未来全文检索的平台。目前，大英图书馆、荷兰视听研究所以及 IA 已经开始探索使用 Solr 解决 WARC 存档项目的全文检索问题。

中国科学院文献情报中心在实施国际重要科研机构网络信息存档项目[10]中，需要构建一个 Web 存档访问服务平台，为科研人员、情报人员、科技管理人员长期有效地利用存档资源提供有利的支撑。这不但需要对存档资源提供基本访问服务能力，同时需要支持用户在时间线上对相应的网络科技信息进行研究和分析，还需要为数据挖掘和再利用提供历史数据的支撑。这些功能的实现需要底层索引系统提供有力的支持，而现有的 WARC 索引工具很难满足这些要求，因此在项目研发过程中，笔者开始探索利用 Solr 平台构建面向 WARC 文件的全文索引系统以满足用户的需求。

在网络资源长期保存与检索领域，与 Solr 相关的研究仍处于起步阶段，存在许多问题需要解决，尤其是 Solr 不具备直接索引 WARC 文件的能力。为了解决上述问题，并从 WARC 文件中提取出更多的信息，本文利用 WARC 文件的结构化特性，采用模块化方案实现 WARC 文件的解析；从解析结果中提取索引字段内容，并利用 Solr 平台构建面向 WARC 文档的全文索引系统。

3 WARC 文件模块化解析的设计与实现

3.1 WARC 文件结构样例及存储方案

WARC 能在单个文件中安全地承载大量数据对象，它提供了将多个资源记录(WARC Record)连接成单个长文件(WARC 文件)的机制，按照互联网上抓取内容块的顺序存储“网络爬取的内容”。为了便于数据的保存与分享，WARC 格式的文件由单个或多个 WARC 记录简单连接而成，第一个记录通常用于描述后续的多个记录。以采集科技网站(http://www.iahe.org/)为例，图 1 给出了采集该站点生成的 WARC 文件所包含的信息。

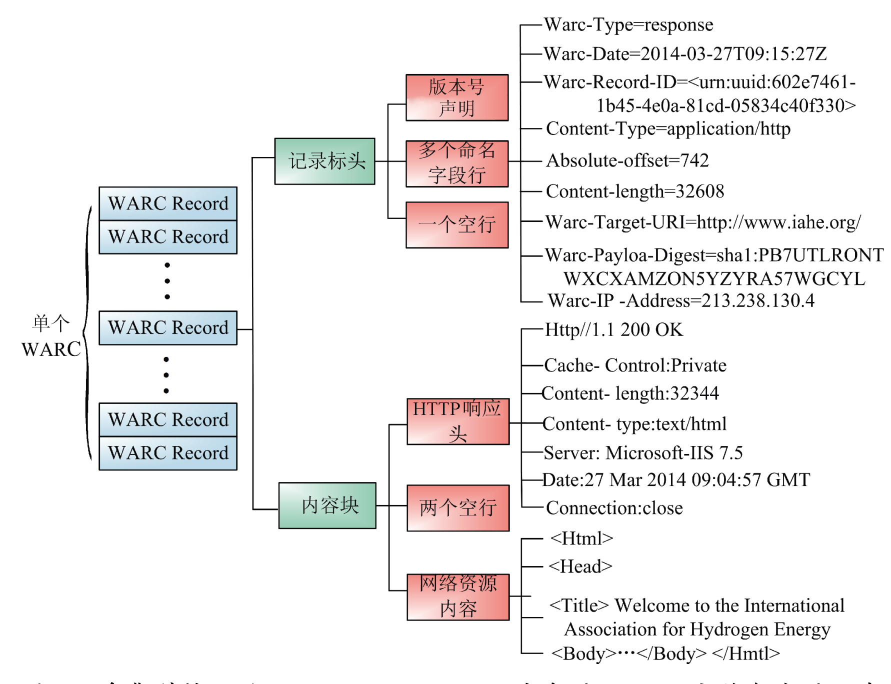


图 1 采集科技网站(http://www.iahe.org/)生成的 WARC 文件包含的信息

一般情况下，WARC 记录的内容或者是一次检索的直接结果(网页、内嵌图片、URL 转向信息、DNS 主机名查询结果、独立文件等)，或者是为存档内容提供附加信息的综合资源(如元数据、转化后的内容)。每个 WARC Record 由一组简单文本标头和任意数据内容块构成，其中 WARC 记录标头的第一行，是用于声明该记录采用给定版本号的 WARC 格式，然后是以空行结束的不定行数的命名字段。可以发现，WARC 文件结构具有分层和结构化的特征。

由于网站内容数据量大，采集网站信息会生成大量 WARC 文件，其储存方案和提供的方式，取决于软件及应用程序的实现。本文使用 Heritrix 对站点进行采集，采集结果可以分由多个 WARC 文件存储，每个 WARC 文件的最大体量可以通过 Heritrix 的配置文件 order.xml 进行设置和管理，即通过写组件<map name="write-processors">中的 max-size-bytes 参数进行设置，并且 Heritrix 会自动对一次采集任务生成的 WARC 文件进行编号，如在采集过程中设置的 WARC 文件最大为 1GB，当对站点采集的数据小于 1GB 时，只生成一个 WARC 文件，WARC 文件名中包含序号 0000，当超过 1GB 时，会分为多个 WARC 文件进行存储，顺序采用 0001, 0002, 0003 等，如图 2 所示：

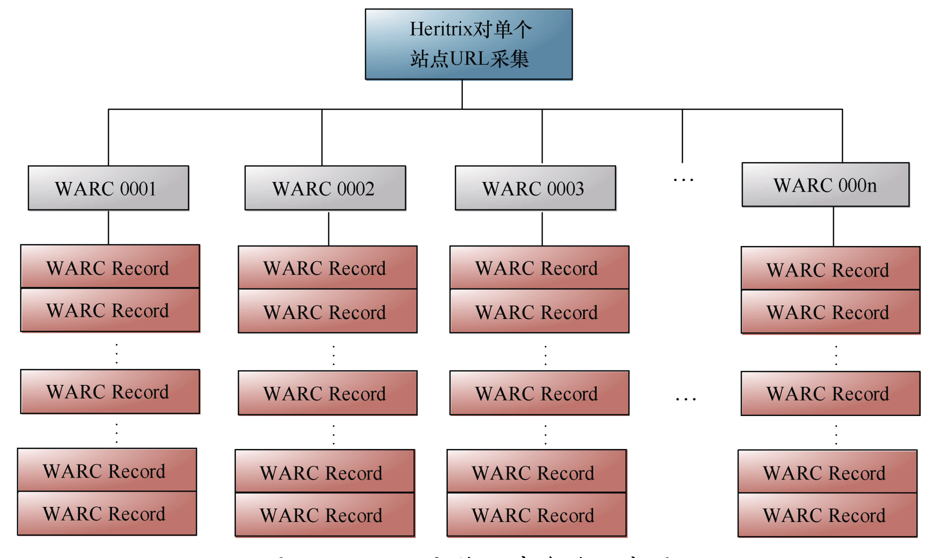


图 2　WARC 文件顺序存储示意图

3.2　WARC 文件模块化解析

根据 WARC 文件的结构化特点，本文在设计 WARC 文件解析方案时采用分层解析以及解析功能模块化的思路。WARC 文件解析流程如图 3 所示，其解析过程可以分为 4 个功能模块：WARC Record 获取模块、标头获取模块、WARC Record 内容块获取模块和内容块解析模块。

(1) WARC Record 获取模块

一个 WARC 文件由多个资源记录 WARC Record 顺序连接构成，通过 WARC Record 获取模块将 WARC 文件拆分为多个 WARC Record，这样就由解析 WARC 文件内容转变为解析 WARC Record 的内容。

Heritrix 源代码中除了网络信息采集的核心工具包之外，还提供针对 WARC 格式的 IO 操作包 org.archive.io.warc。本文利用该包的核心类对 WARC 文件进行读取，如 WARCReader.java，WARCRecord.java, WARCReaderFactory.java 等。在 WARC 文档解析的过程中，循环调用 WARCReaderFactory 类中的 get(String warcFilePath)函数来获取 WARC Reader 对象，遍历整个 WARC 文档，即可以获得每一个 WARC Record。

(2) WARC Record 标头获取模块

每个 WARC Record 结构一致，都包括一组 WARC 记录标头和内容块，在获得 WARC Record 后，可通过 WARC Record 类中的 getHeader()函数完成对记录头标区信息的解析，即获取 WARC Record 标头各命名字段信息，这些信息保存在数据结构 Map 中。对一个 WARC Record 标头的解析结果如表 1 所示。

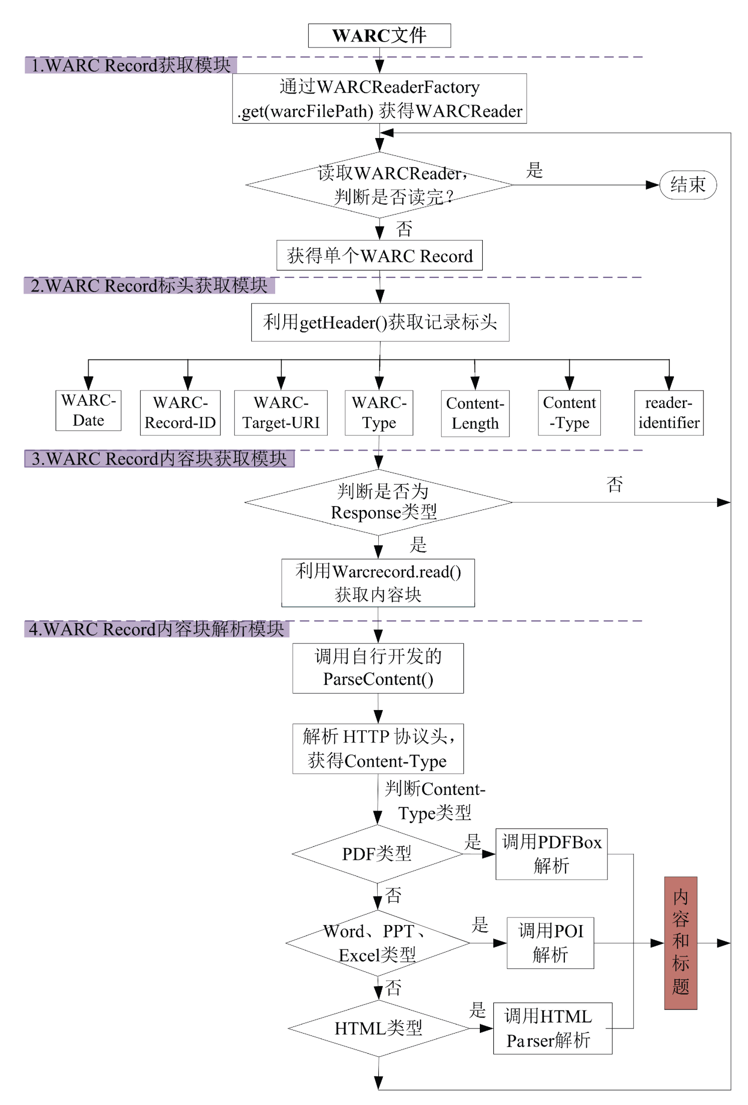


图 3　WARC 文件解析流程

表 1　WARC Record 标头的解析结果样例

序号	Key	Value
1	WARC-Type	response
2	reader-identifier	http://124.16.154.180:8066//warcs/201403/www.iahe.org-20140327091515-00000-hadoop-master-180.warc.gz
3	WARC-Date	2014-03-27T09:15:27Z
4	absolute-offset	742
5	Content-Length	32608
6	WARC-Record-ID	<urn:uuid:602e7461-1b45-4e0a-81cd-05834c40f330>
7	WARC-IP-Address	213.238.130.4
8	WARC-Payload-Digest	sha1:PB7UTLRONTWXCXAMZON5YZYRA57WGCYL
9	WARC-Target-URI	http://www.iahe.org/
10	Content-Type	application/http; msgtype=response

(3) WARC Record 内容块获取模块

类似地，在获得 WARC Record 后，利用 WARC Record 类中的 Warcrecord.read()读取 WARC Record 内容，并以字节数组形式存储。

(4) WARC Record 内容块解析模块

前述模块利用 Heritrix 提供的 WARC 格式 IO 操作包中的核心类获取 WARC Record 标头字段信息和内容块，而对 WARC Record 内容块的解析是通过自行开发模块的 parseContent(byte[] content)完成。解析过程如下：

WARC Record 内容块是由 HTTP 协议头、两个空行和资源内容组成。首先通过空行分隔从 WARC Record 取出 HTTP 协议头部分进行解析，可以获取其

中包含的数据信息。一个 HTTP 协议头的解析结果样例如下:

```
HTTP/1.1 200 OK
Cache-Control: private
Content-Length: 32344
Content-Type: text/html
Server: Microsoft-IIS/7.5
Set-Cookie:ASPSESSIONIDCCCATDCR=BGOBMIKBGEKJJO
 MFKGCDLAJB; path=/
X-Powered-By: ASP.NET
Date: Thu, 27 Mar 2014 09:04:57 GMT
Connection: close
```

该步骤主要目的是通过解析 HTTP 协议头进而获知 Content-Type 字段内容，即明确存档页面的资源类型，然后根据资源类型调用不同的解析工具去解析：若资源类型为 PDF，则调用 PDFBox 解析；若资源类型为 Word、PPT、Excel，则调用 POI 解析；若资源类型为 HTML，则调用 HTMLParser 解析等。在获得存档页面的标题和内容后，将信息保存到数据结构 Map 中。最终，Map 中保存了 WARC Record 的标头字段、标题和内容，完成了对单个 WARC Record 的完整解析。

由于网络资源内容的多样性和复杂性，不排除某些特别资源出现无法解析或解析时间过长的情况，在实际解析过程中，为了保障 WARC 文件的解析效率，笔者定义了一个守护线程去执行 WARC 文件的解析，通过设置守护线程的执行时间来保障解析顺利进行。

从上述过程不难看出，采用模块化解析方案简化了 WARC 文件的解析实现，可以将 WARC 文件内容根据 WARC 结构拆分成相互独立的组成部分，有利于构建内容索引，并可以为以后做数据挖掘、元数据抽取、数据分析服务等提供方便易用的接口。

4 基于 Solr 平台构建 WARC 文件全文索引系统

Web Archive 系统不仅要实现对 WARC 文件的长期保存，而且要能够对 WARC 解析后的内容建立有效的索引，以向用户提供高效的检索和访问功能。考虑到 Solr 在全文索引和分面检索方面的强大优势，本文利用 Solr 工具开发 WARC 文件内容索引模块，以优化用户检索和访问效果，提升用户体验。

4.1 自动索引系统流程设计

为了实现对 WARC 文件的自动化实时索引，开发了一个监测模块，能够实时监测是否有新的 WARC 文件生成；当有新的 WARC 文件生成时，自动将其加入索引队列，等待索引模块进行索引，保障了 WARC 文件索引和检索内容更新的实时性。

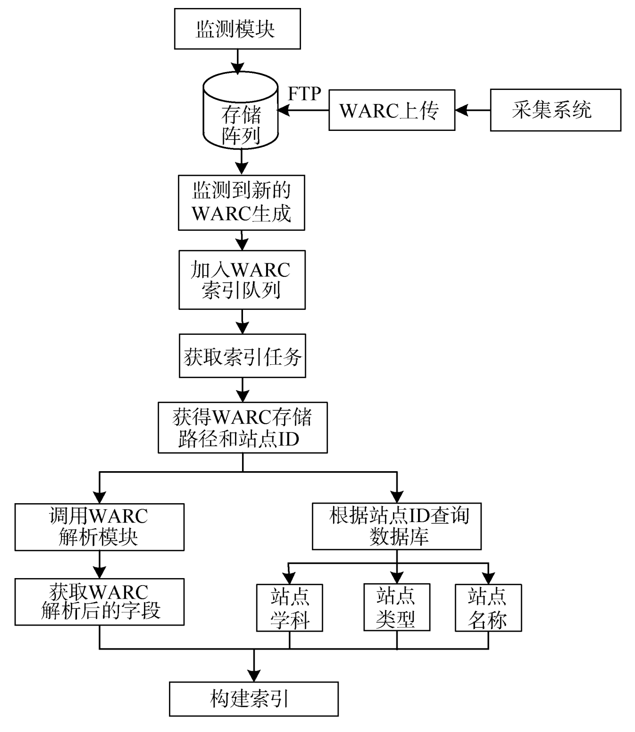


图 4　WARC 文件自动索引系统方案设计

图 4 给出了 WARC 文档自动索引的方案设计，首先要实现 WARC 生成系统、索引模块和 WARC 文件解析模块的对接，达到对 WARC 文件自动化持续索引的目的。中国科学院文献情报中心构建的采集系统对站点进行周期性采集，采集任务完成后自动将生成的 WARC 上传到存储阵列进行统一存储，为了实现对采集生成的 WARC 文件实时索引，监测模块可以周期性对存储阵列进行轮询以查看是否有新的 WARC 生成，当监测到有新的 WARC 文件时，将新的 WARC 文件加入索引队列，等待索引模块实施索引。索引模块从索引队列获取索引任务，然后调用 WARC 获取和解析模块对 WARC 文件进行分层读取和解析，解析结果保存在数据结构 Map 中；索引模块从 Map 中获取要建立的索引字段的值，可以获取的索引字段包括该存档页面的 URL、存档时间、资源类型、存档页面的标题、页面内容等。

4.2 索引字段设计

索引字段的设计除了上述 WARC 解析获得的字段，还包括数据库的辅助扩展字段，如学科分类、来源的站点名称、站点类型等。每条索引对应 WARC 文件中的一个 WARC Record，即一个存档页面的信息。索引字段的详细设计如表 2 所示：

表 2 索引字段设计

序号	字段名称	字段含义	字段来源	备注
1	WARC_Record_ID	页面唯一标识	WARC Record 头部解析	
2	WARC_Target_URI	存档页面 URL	WARC Record 头部解析	
3	WARC_Filename	WARC 文件名	WARC Record 头部解析	
4	WARC_Type	记录类型	WARC Record 头部解析	
5	WARC_Date	存档时间	WARC Record 头部解析	
6	Datelist	存档时间列表	索引查询	
7	year	年	WARC Record 头部解析	设为分面字段
8	month	月	WARC Record 头部解析	
9	day	日	WARC Record 头部解析	
10	type	资源类型	WARC Record 内容解析	
11	formattype	规范资源类型	WARC Record 内容解析	设为分面字段
12	title	存档页面标题	WARC Record 内容解析	
13	content	存档页面内容	WARC Record 内容解析	
14	keyword	页面关键词	WARC Record 内容解析	
15	site_name	站点名称	数据库	设为分面字段
16	site_en_name	站点英文名称	数据库	
17	site_url	站点 URL	数据库	
18	subject	站点学科分类	数据库	设为分面字段

4.3 索引策略设计

为了提高用户的检索效率和更好地揭示存档的 WARC 文件内容，笔者设计了两种不同的索引策略，即 WARC 记录综合索引和存档页面索引。

虽然两种索引策略的每一条记录都代表一个存档 URL，但第一种索引策略是以存档页面的 WARC_Record_ID 为唯一标识符，每一条索引都代表一个完整的 WARC Record，建立索引时需将 WARC 文件的解析字段和数据库读取字段依次写入索引相应字段。第二种索引策略是以存档页面的 URL 为唯一标识符，同时新增一个 datelist 索引字段；该字段为一个多值字段，用于存储该存档页面不同的存档时间，代表该存档页面自存档以来的存档次数和每次存档的时间。在为每一个 WARC Record 建立索引时，首先查询索引中是否存在该 URL 记录；若不存在，则只需将 WARC 文件的解析字段和数据库读取字段依次写入索引相应字段；若索引中已存在该 URL，说明该 URL 已经被存档过，则将此次的存档时间和以前的存档时间合并，重新写入 datelist 字段，而其他字段内容更新为最新存档时间的页面内容。

采取上述两种索引策略的优势在于可以满足用户多样性的查询需求，以 WARC_Record_ID 为唯一标识符的索引可以提供存档 URL 每一次采集的完整信息，以存档页面 URL 为唯一标识符的索引可以将 URL 多次采集信息进行归并处理，能够清晰地展示每一个 URL 的全部采集情况，同时大大减少索引体量，提高检索速度，并且这两种索引可以相互支撑，相互补充。

5 Web Archive 访问平台应用效果分析

中国科学院文献情报中心已基本开发完成面向重要科研机构的 Web Archive 访问平台，实现了 WARC 文件的模块化解析，构建了基于内容的全文索引系统。与常用的开源索引工具相比，该平台具有更完善的访问和检索功能，可以从多角度向用户展示存档资源，如表 3 所示。

表 3　不同访问平台的索引工具功能对比

功能	Wayback	NuchtWAX	WERA	Web Archive 平台
URL 检索	●	○	●	●
内容检索	○	●	●	●
学科分面	○	○	○	●
资源类型分面	○	○	○	●
时间分面	○	○	○	●
站点浏览	○	○	○	●

(注: ●具备功能, ○缺少功能)

截至 2016 年 1 月, Web Archive 平台采集存档数据总量达到了 26TB(压缩包体量), WARC 文件数总计两万多个, 索引体量超过了 500GB; 平台能够提供 URL 和全文检索, 同时增加了学科分类、资源类型、存档时间等分面内容, 实现了从多维度对 WARC 文件内容进行揭示。通过首页内容检索入口输入关键词可以进行内容检索, 如图 5 所示。检索结果给出了采集的 URL, 对文档类型如 HTML、PDF、DOC、PPT、TXT 等提取了标题和内容, 从检索结果中还可以看到每个 URL 的存档次数、所属的学科领域、所属站点名称以及最新存档时间等, 同时还可以对检索结果按相关度和日期进行排序, 如图 6 所示。



图 5　Web Archive 访问平台首页的检索入口及功能分区



图 6　检索结果页面

6　结　语

面向国际重要科技机构网站开展的 Web Archive 系统建设中, 利用开源软件 Heritrix 实现网络内容的大规模分布式采集, 以 WARC 文件格式进行存储, 不仅实现了科技网站资源的长期保存, 也有利于存档资源的可持续发展。基于 Heritrix 提供的 WARC 文件 I/O 操作包, 实现了 WARC 文件的分层读取; 对读取的 WARC 内容, 开展了不同类型网络资源的模块化解析; 奠定了构建 Solr 索引的基础, 同时为进一步的数据挖掘、元数据抽取、数据分析服务等提供相应的接口。

基于 Solr 平台开发的 WARC 文件索引和访问系统, 增加了学科分类、资源类型和存档时间等分面内容, 实现了从多维度对 WARC 文件内容进行揭示。可以提供对存档资源的 URL 和全文检索, 同时明显提高了 WARC 文件内容的检索效率。此外, 该平台能够实时解析和索引生成的 WARC 文件, 跟踪网络资源的动态变化, 对于 Web Archive 的在线实时更新具有重要意义。

Web Archive 资源是一个巨大的宝库, 如何深度挖掘和发挥存档资源的价值, 是未来面对的主要挑战之一, 也是今后研究的方向, 希望笔者的探索能够为网络保存领域的同仁提供有益的参考。

参考文献:

[1] IIPC Members [EB/OL]. [2015-12-25]. http://netpreserve.org/about-us/members.

[2] ISO 28500: 2009 WARC File Format [EB/OL]. [2009-05-15]. http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=44717.

[3] 曲云鹏. 网络存档文件格式 WARC 研究[J]. 图书馆学研究, 2014(24): 20-25, 28. (Qu Yunpeng. Research on the Standardized WARC File Format [J]. Researches on Library Science, 2014(24): 20-25, 28.)

[4] 孙志茹, 吴振新, 曲云鹏. 基于 Wayback 的索引策略研究[J]. 现代图书情报技术, 2009(4): 14-18. (Sun Zhiru, Wu Zhenxin, Qu Yunpeng. Analysis of Index Strategies in Web Archive[J]. New Technology of Library and Information Service, 2009(4): 14-18.)

[5] 吴振新, 曲云鹏, 李成文, 等. 基于开源软件搭建网络信息资源采集与保存平台[J]. 现代图书情报技术, 2009(7-8): 6-10. (Wu Zhenxin, Qu Yunpeng, Li Chengwen, et al.

Constructing a System for Harvesting and Preserving Chinese Web Information Resources Based on Open Source Software [J]. New Technology of Library and Information Service, 2009(7-8): 6-10.)
[6] WERA 0.4.2RC1 [EB/OL]. [2006-01-17]. http://archive-access.sourceforge.net/projects/wera/.
[7] NutchWAX 0.11.0-SNAPSHOT API [EB/OL]. [2007-02-20]. http://archive-access.sourceforge.net/projects/nutchwax/apidocs/overview-summary.html.
[8] SOLR-Nutch Report [EB/OL]. [2011-01-31]. http://archive.org/~aaron/iipc/solr-nutch-report.html.
[9] Solr Features [EB/OL]. [2016-01-25]. http://lucene.apache.org/solr/features.html.
[10] 吴振新, 张智雄, 谢靖, 等. 基于 IIPC 开源软件拓展构建国际重要科研机构 Web 存档系统[J]. 现代图书情报技术, 2015(4): 1-9. (Wu Zhenxin, Zhang Zhixiong, Xie Jing, et al. Developing Web Archive System of International Institutions Based on IIPC Open Source Software [J]. New Technology of Library and Information Service, 2015(4): 1-9.)

作者贡献声明:

胡吉颖: 设计解析和索引方案, 系统开发, 论文撰写;
吴振新: 设计解析和索引方案, 论文撰写及最终版本修订;
谢靖: 设计索引方案, 系统开发;
张智雄: 完善索引方案设计。

利益冲突声明:

所有作者声明不存在利益冲突关系。

收稿日期: 2016-02-25
收修改稿日期: 2016-03-22

A Full-text Indexing System for WARC Files

Hu Jiying　Wu Zhenxin　Xie Jing　Zhang Zhixiong
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] This paper develops a parsing and indexing system for the WARC files, which fully exploits the value of Web archives from scientific institutions. [**Context**] The WARC files have been widely used in digital curation. However, the existing full-text indexing tools cannot satisfy the diversified needs of the WARC searchers. [**Methods**] We employed a modular scheme to parse the WARC files. Upon analyzing popular indexing tools, developed a new full-text indexing system based on the Solr platform. [**Results**] The new system effectively indexed the Web archives. Users could search information from different perspective, such as the subject category, resource type, and archived time, etc. [**Conclusions**] The new system indexes rich Web archives from international institutions, and improves the efficiency of users' information retrieval activities.

Keywords: Web archive　WARC file　Modular parse　Solr index